



# Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis

## Citation

Namjou, B., K. Marsolo, R. J. Carroll, J. C. Denny, M. D. Ritchie, S. S. Verma, T. Lingren, et al. 2014. "Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis." *Frontiers in Genetics* 5 (1): 401. doi:10.3389/fgene.2014.00401. <http://dx.doi.org/10.3389/fgene.2014.00401>.

## Published Version

doi:10.3389/fgene.2014.00401

## Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:13581255>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)



# Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links *PLCL1* to speech language development and *IL5-IL13* to Eosinophilic Esophagitis

**Bahram Namjou<sup>1,2\*</sup>, Keith Marsolo<sup>2,3</sup>, Robert J. Carroll<sup>4</sup>, Joshua C. Denny<sup>4,5</sup>, Marylyn D. Ritchie<sup>6</sup>, Shefali S. Verma<sup>6</sup>, Todd Lingren<sup>2,3</sup>, Aleksey Porollo<sup>1,2,3</sup>, Beth L. Cobb<sup>1</sup>, Cassandra Perry<sup>7</sup>, Leah C. Kottyan<sup>1,2,8</sup>, Marc E. Rothenberg<sup>8</sup>, Susan D. Thompson<sup>1,2</sup>, Ingrid A. Holm<sup>9</sup>, Isaac S. Kohane<sup>10</sup> and John B. Harley<sup>1,2,11</sup>**

<sup>1</sup> Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

<sup>2</sup> College of Medicine, University of Cincinnati, Cincinnati, OH, USA

<sup>3</sup> Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

<sup>4</sup> Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA

<sup>5</sup> Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA

<sup>6</sup> Center for Systems Genomics, The Pennsylvania State University, Philadelphia, PA, USA

<sup>7</sup> Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA

<sup>8</sup> Division of Allergy and Immunology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

<sup>9</sup> Division of Genetics and Genomics, Department of Pediatrics, The Manton Center for Orphan Disease Research, Harvard Medical School, Boston Children's Hospital, Boston, MA, USA

<sup>10</sup> Children's Hospital Informatics Program, Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA

<sup>11</sup> U.S. Department of Veterans Affairs Medical Center, Cincinnati, OH, USA

## Edited by:

Mariza De Andrade, Mayo Clinic, USA

## Reviewed by:

Andrew Skol, University of Chicago, USA

Albert Vernon Smith, Icelandic Heart Association, Iceland

Shelley Cole, Texas Biomedical Research Institute, USA

## \*Correspondence:

Bahram Namjou, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA  
e-mail: bahram.namjou@cchmc.org

**Objective:** We report the first pediatric specific Phenome-Wide Association Study (PheWAS) using electronic medical records (EMRs). Given the early success of PheWAS in adult populations, we investigated the feasibility of this approach in pediatric cohorts in which associations between a previously known genetic variant and a wide range of clinical or physiological traits were evaluated. Although computationally intensive, this approach has potential to reveal disease mechanistic relationships between a variant and a network of phenotypes.

**Method:** Data on 5049 samples of European ancestry were obtained from the EMRs of two large academic centers in five different genotyped cohorts. Recently, these samples have undergone whole genome imputation. After standard quality controls, removing missing data and outliers based on principal components analyses (PCA), 4268 samples were used for the PheWAS study. We scanned for associations between 2476 single-nucleotide polymorphisms (SNP) with available genotyping data from previously published GWAS studies and 539 EMR-derived phenotypes. The false discovery rate was calculated and, for any new PheWAS findings, a permutation approach (with up to 1,000,000 trials) was implemented.

**Results:** This PheWAS found a variety of common variants (MAF > 10%) with prior GWAS associations in our pediatric cohorts including Juvenile Rheumatoid Arthritis (JRA), Asthma, Autism and Pervasive Developmental Disorder (PDD) and Type 1 Diabetes with a false discovery rate < 0.05 and power of study above 80%. In addition, several new PheWAS findings were identified including a cluster of association near the *NDFIP1* gene for mental retardation (best SNP rs10057309,  $p = 4.33 \times 10^{-7}$ ,  $OR = 1.70$ , 95%CI = 1.38 – 2.09); association near *PLCL1* gene for developmental delays and speech disorder [best SNP rs1595825,  $p = 1.13 \times 10^{-8}$ ,  $OR = 0.65(0.57 - 0.76)$ ]; a cluster of associations in the *IL5-IL13* region with Eosinophilic Esophagitis (EoE) [best at rs12653750,  $p = 3.03 \times 10^{-9}$ ,  $OR = 1.73$  95%CI = (1.44 – 2.07)], previously implicated in asthma, allergy, and eosinophilia; and association of variants in *GCKR* and *JAZF1* with allergic rhinitis in our pediatric cohorts [best SNP rs780093,  $p = 2.18 \times 10^{-5}$ ,  $OR = 1.39$ , 95%CI = (1.19 – 1.61)], previously demonstrated in metabolic disease and diabetes in adults.

**Conclusion:** The PheWAS approach with re-mapping ICD-9 structured codes for our European-origin pediatric cohorts, as with the previous adult studies, finds many previously reported associations as well as presents the discovery of associations with potentially important clinical implications.

**Keywords:** PheWAS, ICD-9 code, genetic polymorphism

## INTRODUCTION

Phenome-wide association study (PheWAS) is a relatively new genomic approach to link clinical conditions with published variants (Denny et al., 2010). The concept, although not new, was originally applied to genomic research by the eMERGE (electronic MEDical Records and GENomics) network, which is in a unique position to access tens of thousands of Electronic Medical Records (EMR) linked to ICD-9 codes in structured data. Multiple eMERGE PheWAS results have been published that primarily address adult cohorts (Denny et al., 2011, 2013). The phenotypic data used in PheWAS may include ICD-9 codes, epidemiologic data in health surveys, biomarkers, intermediate or quantitative traits (Pendergrass et al., 2011, 2013; Neuraz et al., 2013; Liao et al., 2014). By virtue of this inclusive approach, new hypotheses may be generated that provide insight into genetic architecture of complex traits. Challenges with PheWAS include multiple test corrections across the thousands of phenotypes tested and autocorrelation of some of the phenotypes. Nevertheless, novel robust insights have resulted from PheWAS, for example, genetic association findings with heart rate variability are notable (Ritchie et al., 2013).

PheWAS combines multiple phenotypes from previous GWAS, and identify common SNPs affecting different traits. In this study, we used this approach to evaluate whether known GWAS variants identified in adult diseases can be also identified in children using two EMR-linked pediatric datasets from eMERGE. PheWAS in pediatrics is particularly important because it not only assesses the effect of early age of onset on many established adult-GWAS loci, but also may provide insights into how a primary phenotype during child development develops into one or more diseases in adulthood. A priori, there are several reasons that in principle might make a pediatric PheWAS more challenging. These include the change in heritability with age for several traits (St Pourcain et al., 2014), the flux in the recommendations for pediatric monitoring for traits that are routinely measured in adults (Gidding, 1993; Klein et al., 2010) and the use of cross-sectional standardization rather than longitudinal standardization of developmental traits such as height (Tiisala and Kantero, 1971).

To determine whether robust association signals would be present in the context of these challenges, we conducted the first PheWAS study in pediatrics on our available samples. We successfully translated 93,724 specific ICD-9 diagnostic codes into 1402 distinct PheWAS code groups and 14 major disease concept paths and evaluated 2481 previously published variants. After quality control, only 2476 genetic variants were analyzed in 539 diseases in the two pediatric sites. Finally we replicated 24 genetic variants and identified 14 new possible associations confirming our hypothesis. Our primary results highlight the utility of an EMR-based PheWAS approach as a new line of investigation for discovery of genotype-phenotype associations in pediatrics.

## MATERIALS AND METHODS

### STUDY SUBJECTS

Protocols for this study were approved by the Institutional Review Boards (IRBs) at the institutions where participants were recruited. All study participants provided written consent prior to study enrolment; consent forms were obtained at

each location under IRB guidelines. Children and teens, aged through 19 years old were included. The EMR-linked pediatric emerge cohorts consist of 4560 subjects from Cincinnati Children's Hospital Medical Center (CCHMC) and 1000 subjects from Boston Children's Hospital (BCH). Only those self-reported to have European ancestry were selected for this study (Table 1).

### SNP PRIORITIZATION

We limit our investigation to particular genetic variants: First, we obtained the list of all previously published SNPs from different public domain databases including The National Human Genome Research Institute (NHGRI) catalog of published Genome-Wide Association Studies (<http://www.genome.gov/gwastudies>), Genetic Association of Complex Diseases and Disorders (GAD, <http://geneticassociationdb.nih.gov>), the UCSC Genome Browser database (UCSC, <http://genome.ucsc.edu/>), Online Mendelian Inheritance in Man (OMIM, <http://www.omim.org/>), and PharmGKB (<http://www.pharmgkb.org>). After linking this collection to PubMed reference numbers, only those with at least one reported of positive associations were selected regardless of the previously observed *p* values or number of publications. In addition, all downloaded databases were current at the time of this submission. From the filtered variants, 2476 variants were available and assessed in our clean, post-imputation genotyping dataset for analysis.

### GENOTYPING AND STATISTICAL ANALYSES

High throughput SNP genotyping was carried out previously in CCHMC and BCH using Illumina™ or Affymetrix™ platforms, as previously described (Namjou et al., 2013). Quality control (QC) of the data was performed before imputation. In each genotyped cohort, standard quality control criteria were met and single nucleotide polymorphisms (SNPs) were removed if (a) >5% of the genotyping data was missing, (b) out of Hardy-Weinberg equilibrium (HWE,  $p < 0.001$ ) in controls, or a minor allele frequency (MAF) <1%. Samples with call rate <98% were excluded.

Recently all eMERGE cohorts have also undergone whole genome imputation. The details of these procedures are available in this issue of *Frontiers in Genetics* (Setia et al., 2014). Briefly, the imputation pipeline was implemented using IMPUTE2 program and the publicly available 1000-Genomes Project as the reference haplotype panel composed of 1092 samples (release version 2 from March 2012 of the 1000 Genomes Project Phase I, <ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521>) (Howie et al., 2011). The eMERGE imputed data provided to us were already filtered, i.e., imputed data with a threshold of 0.90 for the genotype posterior probability and with a IMPUTE2 info score > 0.7 (Howie et al., 2011). Principle component analysis (PCA) performed to identify outliers and hidden population structure using EIGENSTRAT (Price et al., 2006). The first two principle components explained most of the variance and were retained and used as covariates during the association analysis in order to adjust for population stratification. In addition, 14 outlier samples were removed. To illustrate the overall inflation rate a phenotype with sufficient number of cases and

**Table 1 | The demographic distribution of the European ancestry population (CCHMC-BCH).**

	Cohort names	#Europeans	M/F	Mean age (95%CI)	Array
BCH*	The gene partnership	727	449/278	13.30(12.97–13.66)	Affymetrix-Axiom
CCHMC**	Cytogenetics	1228	758/470	7.32(7.03–7.62)	Illumina-610
	Cytogenetics	609	373/236	7.18(6.73–7.63)	Illumina-Omni-1
	EoE <sup>†</sup>	543	394/149	12.27 (11.70–12.67)	Illumina-Omni-5
	JIA <sup>‡</sup>	488	101/387	13.70(13.13–14.23)	Affymetrix-6
	Cincinnati- control cohorts	673	329/344	13.50(13.25–13.84)	Illumina-Omni-5
Total		4268	2403/1865	11.52(11.16–11.91)	

\*BCH, Boston Children's Hospital; \*\*CCHMC, Cincinnati Children's Hospital Medical center; †, Eosinophilic Esophagitis (EoE) cohorts; ‡, Juvenile Idiopathic Arthritis cohorts (JIA). The details of platforms used have been described elsewhere (Namjou et al., 2013).

controls has been selected (autism) and the inflation of  $\lambda = 1.03$  was obtained.

Next, from our prioritized SNP list mentioned above, 2481 variants were available. Five of these SNPs had a site-specific effect with either CCHMC or BCH ( $p < 10^{-5}$  for the difference between sites) and were removed from final analyses. For each phenotype, logistic regression was performed between cases and control adjusted for two principal components using PLINK (Purcell et al., 2007). To investigate whether either the phenotype or the genotype has an effect on the outcome variable, we perform phenotypic and genotypic conditional analyses, controlling for the effect of a specific SNP or phenotype. After pruning of highly correlated SNPs ( $r^2 > 0.5$ ), we used false discovery rate (FDR) methods to correct for multiple testing using the Benjamini–Hochberg procedure implemented in PLINK (Purcell et al., 2007). As a result of LD pruning 1828 independent variants were used for the purpose of FDR estimation. Q values correspond to the proportion of false positives among the results. Thus, Q values less than 0.05 signify less than 5% of false positives and are accepted as a measure of significance ( $FDR < 0.05$ ) in this study. For any novel PheWAS findings, an adaptive permutation approach was performed using a sample randomization strategy in which case and control labels were permuted randomly (with up to 1,000,000 trials) in order to obtain empirical  $p$  values [PLINK (Purcell et al., 2007)]. We also report previous known effects that only produce suggestive findings in our study ( $0.05 < p < 0.001$ ). Sample size and power calculations based on the size effect and risk allele frequency were estimated using QUANTO (Gauderman and Morrison, 2006). To graphically display results, LocusZoom was used (Pruim et al., 2010).

## PHENOTYPING

A phenome-wide association analysis (PheWAS) was performed in which presence or absence of each PheWAS code [mapped from translated ICD-9 codes as per Carroll et al., 2014)] were considered as a binary phenotype. The per-patient ICD-9 codes were obtained from the i2b2 Research Patient Data Warehouse at CCHMC and BCH. Also, these PheWAS codes were used to define comparison control groups by excluding the PheWAS case- code and those closely related to them in the ICD-9 hierarchy. Control groups for Crohn's Disease (CD), for instance,

excluded CD, ulcerative colitis, and several other related gastrointestinal complaints. Similarly, control groups for myocardial infarction excluded patients with myocardial infarctions, as well as angina and other evidence of ischemic heart disease. The current PheWAS map and PheWAS script written in R is available [http://phewascatalog.org, (Carroll et al., 2014)]. In this study, subgroups of European cases with more than 20 samples were selected for PheWAS association study (539 subgroups) and the available published SNPs that passed quality controls were evaluated. The case cohorts for the two phenotypes of Juvenile Idiopathic Arthritis (JIA) and Eosinophilic Esophagitis (EoE) have both been previously published as parts of larger phenotype specific studies (Rothenberg et al., 2010; Thompson et al., 2012; Hinks et al., 2013). The origin of all case records is presented in **Table 1**. In this study, Juvenile Onset Rheumatoid Arthritis (JRA) is identified by ICD-9 codes and designated as JRA; when the criteria for Juvenile Idiopathic Arthritis (JIA) were applied in the studies of others (Thompson et al., 2012), then this phenotype was referred to as JIA.

## RESULTS

In this study only European ancestry was included in the analysis to avoid potential bias induced by ancestry. The demographic distribution of the European ancestry population under study (**Table 2**) had 93,724 specific ICD-9 diagnostic codes representing 1402 distinct PheWAS code groups and 14 major disease concept paths. The frequencies of concept path hierarchy of the ontology (**Figure 1**) show the neuropsychiatric concept path as the most frequent and neoplastic and infection paths as the least frequent.

### Replication of existing associations using PheWAS

We compared SNPs with previous GWAS-reports and present association findings ( $FDR-q < 0.05$ ) after corrected for population stratification and standard quality control (**Table 2**).

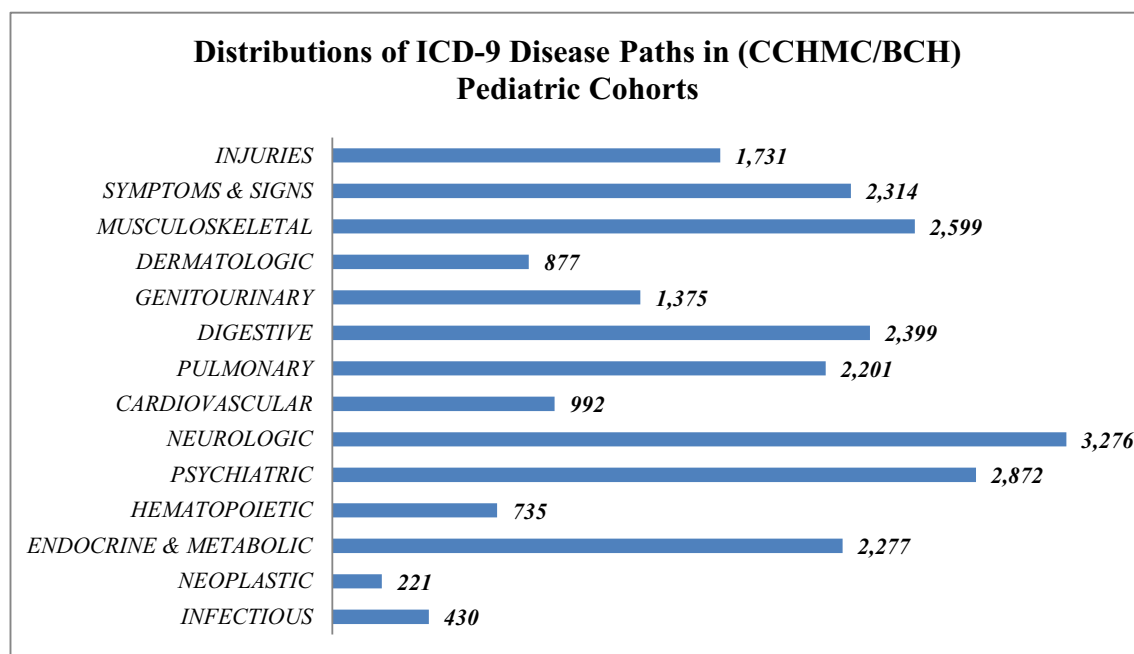
First, for the two phenotypes of JRA and EoE samples overlap largely with those previously reported phenotype specific GWAS study (Rothenberg et al., 2010; Thompson et al., 2012; Kottyan et al., 2014). We reproduced the major findings of those publications using different methodology. For JRA, association with *PTPN22* is a consistent finding. As expected, we replicated a previous report of association of *PTPN22* at non-synonymous coding SNP rs2476601 with this phenotype and with the same direction

**Table 2 | Replication of previous GWAS association results in CCHMC/BCH pediatric cohorts.**

Chr	SNP	Position	Gene	Minor allele	Case	Control	p value	FDRq value	OR	Description	Case/Control
1	rs2476601	114377568	PTPN22	A	0.16	0.09	9.10E-07	8.01E-06	1.87 (1.46–2.41)	JRA	272/3412
1	rs2476601	114377568	PTPN22	A	0.28	0.10	2.78E-05	4.16E-04	3.44 (1.80–6.57)	Thyroiditis	23/3571
1	rs2476601	114377568	PTPN22	A	0.18	0.10	0.007	NS	1.96 (1.16–3.31)	T1DM	47/3609
1	rs6679677	114303808	PTPN22	A	0.16	0.09	3.63E-07	4.15E-06	1.92 (1.49–2.47)	JRA	272/3412
1	rs6679677	114303808	PTPN22	A	0.28	0.10	2.00E-05	4.16E-04	3.52 (1.84–6.74)	Thyroiditis	23/3571
1	rs6679677	114303808	PTPN22	A	0.18	0.10	0.005	NS	2.00 (1.18–3.38)	T1DM	47/3609
2	rs3771180	102953617	IL1RL1	T	0.19	0.14	5.71E-05	0.0005	1.46 (1.19–1.80)	EoE or Food Allergy	599/2346
2	rs7574865	191964633	STAT4	T	0.32	0.24	0.004	NS	1.46 (1.11–1.92)	Wheezing	125/3372
3	rs78122814	85200034	CADM2	A	0.08	0.05	4.34E-05	0.0004	1.72 (1.32–2.24)	Autism	601/1840
5	rs3806932	110405675	TSLP	G	0.35	0.44	5.59E-07	8.38E-06	0.69 (0.59–0.80)	EoE	446/2586
5	rs272889	131665378	SLC22A4	A	0.46	0.37	1.53E-05	0.0003	1.45 (1.22–1.71)	Atopic Dermatitis	298/3031
5	rs12653750	131971902	IL5-IL13	T	0.27	0.20	9.74E-05	0.0005	1.50 (1.22–1.84)	Eosinophilia	250/3344
6	rs75732170	101845494	GRIK2	A	0.06	0.03	8.49E-06	0.0002	2.00 (1.47–2.73)	Autism	601/1840
6	rs477515	32569691	HLA-DRB1	A	0.17	0.33	1.15E-12	8.62E-12	0.41 (0.32–0.53)	JRA	272/3412
6	rs477515	32569691	HLA-DRB1	A	0.07	0.33	1.12E-06	2.60E-05	0.16 (0.08–0.38)	Uveitis	51/3089
6	rs622137	32569852	HLA-DRB1	A	0.17	0.32	4.98E-13	5.78E-12	0.41 (0.32–0.53)	JRA	272/3412
6	rs2516051	32570184	HLA-DRB1	T	0.17	0.32	5.78E-13	5.78E-12	0.41 (0.32–0.53)	JRA	272/3412
6	rs2516049	32570400	HLA-DRB1	C	0.14	0.32	1.49E-15	4.48E-14	0.36 (0.27–0.46)	JRA	272/3412
6	rs660895	32577380	HLA-DRB1	G	0.42	0.21	7.85E-07	1.65E-05	2.73 (1.80–4.13)	T1DM	47/3609
6	rs9388489	126698719	CENPW	G	0.68	0.47	3.07E-05	0.0003	2.46 (1.58–3.80)	T1DM	47/3609
6	rs1490388	126835655	CENPW	T	0.68	0.47	4.29E-05	0.0003	2.42 (1.56–3.74)	T1DM	47/3609
9	rs7850258	100549013	FOXE1	A	0.15	0.34	0.005	NS	0.35 (0.15–0.78)	Thyroiditis	23/3571
9	rs1443438	100550028	FOXE1	T	0.15	0.34	0.009	NS	0.35 (0.15–0.78)	Thyroiditis	23/3571
10	rs12411988	65315397	REEP3	C	0.20	0.14	9.50E-05	0.005	1.53 (1.23–1.92)	JRA	272/3412
10	rs7903146	114758349	TCF7L2	T	0.44	0.29	0.001	NS	2.00 (1.29–3.08)	Abnormal Glucose Test	42/3609
16	rs12924729	11187783	CLEC16A	A	0.26	0.35	3.34E-08	9.08E-06	0.67 (0.58–0.77)	EoE or Food Allergy	599/2346
17	rs8067378	38051348	GSDMB	A	0.57	0.49	3.13E-06	0.0001	1.37 (1.19–1.57)	Asthma	499/3175
17	rs2290400	38066240	GSDMB	C	0.43	0.50	1.05E-05	0.0002	0.74 (0.64–0.84)	Asthma	499/3175
17	rs8074094	45348021	ITGB3	C	0.30	0.25	2.00E-05	0.0002	1.29 (1.15–1.45)	PDD	1141/1840
20	rs716316	14908741	MACROD2	T	0.32	0.39	2.01E-05	0.0003	0.74 (0.65–0.85)	Autism	601/1840

False discovery rate (FDR-q < 0.05) was set for the threshold of significance. The calculated odds ratio was based on minor allele frequency and the coded alleles were shown. All positions were based on NCBI build 37. NS (not significant). The p-values and q-values are ordered based on chromosome and position.





**FIGURE 1 |** Frequency and distribution of 14 major ontology concept path categories from CCHMC/BCH European pediatric cohorts.

of allele frequency, ( $p = 9.10 \times 10^{-7}$ ,  $OR = 1.87$ , 95%CI 1.46 – 2.40). The SNP in proxy (rs6679677,  $r^2 = 1$ ) also produced a similar result (**Table 2**). In our cohorts, variants in *PTPN22* are also associated with thyroiditis as well as Type 1 diabetes mellitus (T1DM), consistent with previous reports and despite low sample size (**Table 2**) (Plenge et al., 2007; Todd et al., 2007; Lee et al., 2011). From these three known associations of *PTPN22*, i.e., JRA, T1DM, and thyroiditis, the largest magnitude of the association is with pediatric onset thyroiditis (**Table 2**,  $OR = 3.52$  95%CI 1.84 – 6.75).

For JRA, multiple loci in the HLA region were also associated at the level of  $p < 10^{-12}$  including rs477515 and rs2516049 near *HLA-DRB1* (**Table 2**). Of note, the size effect of HLA related SNPs, were highest for those with coexisting uveitis (best SNP rs477515,  $OR = 6.5$ , 95% CI = 2.73 – 15.68 for the risk allele, **Table 2**). In addition, for JRA, another previously published association (rs12411988 in *REEP3*) was also found and with the same size effect as previously described ( $OR = 1.53$ ) (**Table 2**) (Thompson et al., 2012).

Furthermore, with regard to EoE traits, we also replicated previous major finding of association of SNP rs3806932 located at the vicinity of the *TSLP* gene at 5q22 region [ $p = 5.59 \times 10^{-7}$ ,  $OR = 0.69$  (95%CI = 0.59 – 0.80)] in these cohorts (**Table 2**) (Rothenberg et al., 2010; Kottyan et al., 2014).

For asthma, the best PheWAS results were detected at 17q21 which includes *GSDMB* and has been previously reported to be associated specifically with childhood onset Asthma (Verlaan et al., 2009). In fact, the best associated SNP rs8067378 in our cohorts [ $p = 3.13 \times 10^{-6}$ ,  $OR = 1.37$  (1.19 – 1.57)], tags the asthma associated haplotype in which the allele-specific expression analyses for this haplotype has previously shown strong

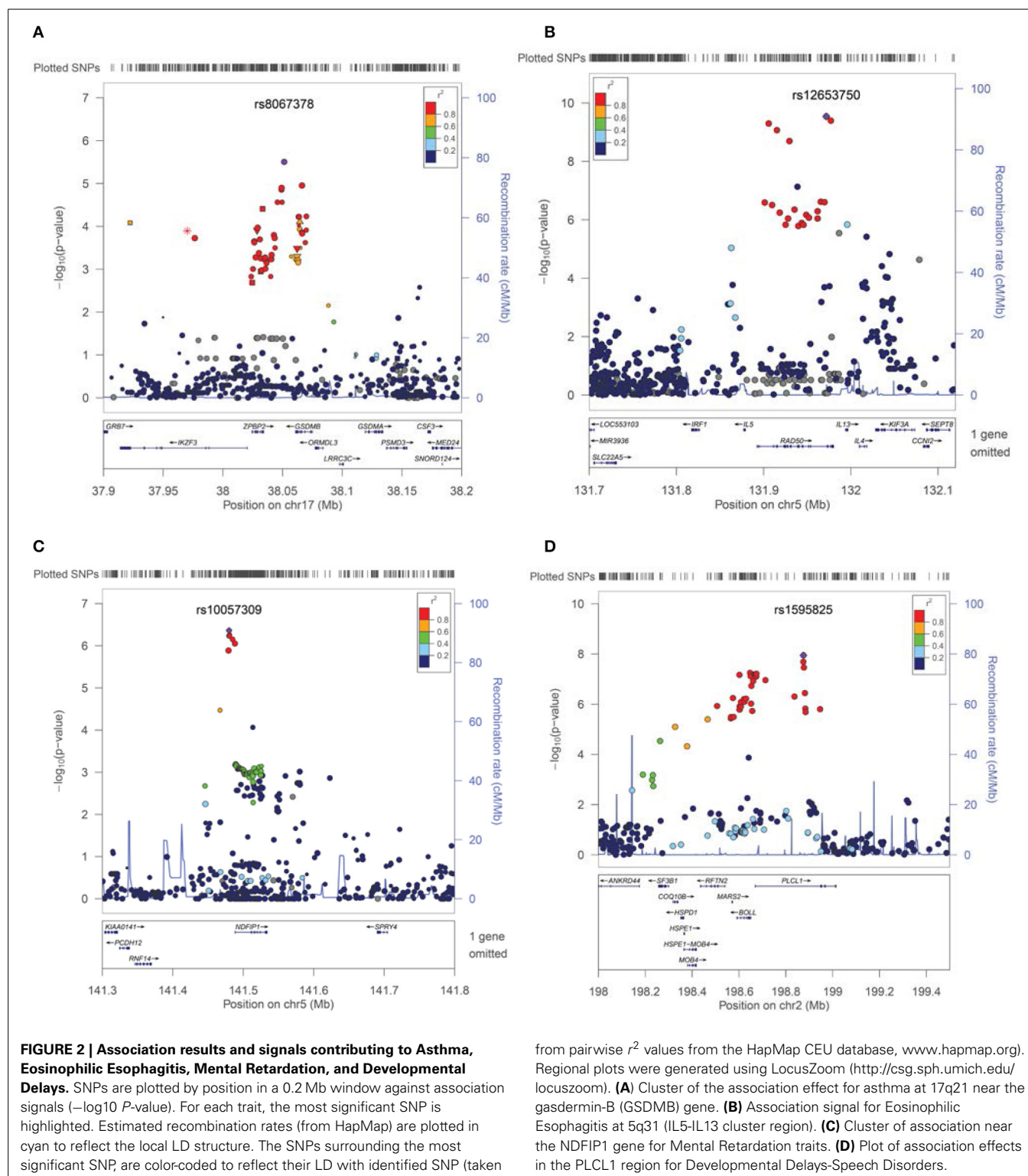
association with Asthma risk (Verlaan et al., 2009). There is strong support for this association from a cluster of variants in this neighborhood (**Figure 2A**).

The minor allele (T) of the intronic SNP rs7903146 in *TCF7L2* is one of the larger magnitude and more frequently identified associations in Type 2 diabetes mellitus (T2DM) and hyperlipidemia in many adult GWAS studies (Lyssenko et al., 2007; Huertas-Vazquez et al., 2008). In fact, the best PheWAS trait in our cohorts at this variant was also related to T2DM and hyperlipidemia as well, although our sample size was small. In this family of ICD-9 codes the best suggestive result was obtained for an abnormal glucose test with [ $p = 0.001$ ,  $OR = 2.00$  (95%CI 1.29 – 3.08)] (**Table 2**).

Specifically, for T1DM, in addition to the positive association with *PTPN22* mentioned above, additional published loci were confirmed and with relatively larger effect sizes ( $OR > 2$ ) including known HLA-SNP rs660895 [ $p = 7.85 \times 10^{-7}$ ,  $OR = 2.73$  (95%CI = 1.80 – 4.13)], as well as variants near *CENPW* that previously have been reported for this trait (**Table 2**) (Barrett et al., 2009).

#### Other effects

Several loci previously associated with autism and pervasive developmental disorders (PDD) (GWAS or copy number variations reports) including those at *MACROD2*, *ITGB3*, *CADM2*, and *GRIK2* (Jamain et al., 2002; Weiss et al., 2006; Thomas et al., 2008; Anney et al., 2010) also provided evidence of association in our cohorts for these traits (**Table 2**). Variants in the *FOXE1* gene that have been previously associated with primary hypothyroidism and thyroiditis in adult eMERGE cohorts (Denny et al., 2011), produced a trend of association and consistent in



directionality with thyroiditis in our pediatric cohorts despite low sample size (Table 2). No gene-gene interaction was evident between *PTPN22* and *FOXE1* for hypothyroidism in these data. Rs7574865 is a SNP in the third intron of the *STAT4* that has been associated with SLE and related autoimmune diseases (Namjou

et al., 2009). In these cohorts, pediatric onset lupus was under-represented (less than 20 cases), however, suggestive associations with wheeze and asthma were detected [ $p = 0.004$ , OR 1.46 (95%CI = 1.11 – 1.92) (Table 2)] with the same direction of the difference in allele frequency previously observed in autoimmune

traits. This possible association has also been reported in another study (Pykäläinen et al., 2005). Of note, in contrast to rheumatoid arthritis, the *STAT4* association effect was weak for JRA in our cohorts (effect size = 1.12,  $p = 0.17$ ). GWAS studies have linked Inflammatory Bowel disease (IBD) to a number of IL-23 pathway genes, in particular *IL23R*. The well-known coding variant in the IL23 receptor (rs11209026) also showed a trend toward association with IBD in our cohorts with the same allelic direction but due to low sample size (31 cases) it did not reach significance (FDR- $q > 0.05$ ) (Li et al., 2010) (data not shown).

### Novel findings from this PheWAS

A number of potentially novel associations remained significant after the permutation procedure to assess the probability of the observed distribution with  $\beta > 0.8$  FDR- $q < 0.05$  (Table 3). Variants in the Glucokinase Regulator gene (*GCKR*) have been previously implicated in metabolic disease, diabetes and hypertriglyceridemia in adults (Bi et al., 2010; Onuma et al., 2010) and were mostly associated with allergic rhinitis in our pediatric cohorts [best SNP rs780093  $p = 2.18 \times 10^{-5}$ ,  $p_{\text{(perm)}} = 8.06 \times 10^{-5}$ , OR = 1.39, 95%CI = (1.19 – 1.61)] (Table 3), while no significant association was found for diabetes. Indeed, conditional analyses, controlling for diabetes related traits suggest that this is an independent effect ( $p_{\text{conditional}} = 6.75 \times 10^{-5}$ ). Another major regulatory locus for diabetes in adults, *JAZF1*, also was associated with allergic rhinitis in our cohorts (Table 3) even after controlling for diabetes ( $p_{\text{conditional}} = 8.46 \times 10^{-5}$ , for rs1635852). No significant gene-gene interaction was detected between these two loci or with *TCF7L2*.

Variants in a cytokine cluster of the *IL5-IL13* region, which is known to be associated with Asthma, Allergy, Atopic Dermatitis (AD) and Eosinophilia, produced a cluster of association with EoE in our cohorts [best SNP rs12653750,  $p = 3.03 \times 10^{-9}$ ,  $p_{\text{(perm)}} = 1.00 \times 10^{-6}$ , OR = 1.73 (1.44 – 2.07)] (Bottema et al., 2008; Granada et al., 2012). There is a cluster of significant variants in this neighborhood of chromosome 5 (5q31) associated with EoE (Figure 2B). In our cohorts, weaker associations can be detected for all allergy-related phenotypes with the association with Eosinophilia being the most impressive [ $p = 9.74 \times 10^{-5}$  (Table 2)]. However, conditional analyses and controlling for Asthma and Eosinophilia suggest that an independent effect still exists for EoE at this locus using EMR data (conditional  $p = 9.74 \times 10^{-5}$  for rs20541). Moreover, no long distance linkage disequilibrium between rs3806932 in *TSLP* gene at 5q22 and rs20541 was detected in this population ( $r^2 = 0.0002$ ,  $D' = 0.02$ ).

We also observed association with AD within this cytokine cluster consistent with previous reports (Paternoster et al., 2011). However, the best associated SNP for AD (rs272889) was located at *SLC22A4* in our population (Table 2). These two variants, rs272889 and rs12653750, were separated by more than 300kb with low linkage disequilibrium ( $r^2 < 0.1$ ). A residual effect still exists for AD and rs272889 after controlling for EoE status or the rs12653750 variant that suggests a distinct effect ( $p_{\text{conditional}} = 0.002$ ). Noteworthy, with regard to AD, another reported SNP (rs2897442) downstream of this cluster at *KIF3A* gene produced only a suggestive association ( $p = 0.005$ ) in our cohort (data not shown).

Table 3 | Novel PheWAS findings in CCHMC/BCH pediatric cohorts.

Description	Case/Control	Chr	SNP	Position	Gene	Minor allele	Case	Control	p value	p-permute	Case needed*	OR
Allergic rhinitis	408/2754	2	rs1260326	27730940	GCKR	T	0.48	0.41	702E-05	1.21E-04	250	1.36 (1.17–1.58)
Allergic rhinitis	408/2754	2	rs780094	27741237	GCKR	T	0.47	0.40	2.94E-05	9.61E-05	250	1.38 (1.19–1.60)
Allergic rhinitis	408/2754	2	rs780093	27742603	GCKR	T	0.47	0.40	2.18E-05	8.06E-05	250	1.39 (1.19–1.61)
Allergic rhinitis	408/2754	7	rs864745	28180556	JAZF1	C	0.43	0.50	9.02E-05	1.11E-04	220	0.76 (0.65–0.88)
Allergic rhinitis	408/2754	7	rs1635852	28189411	JAZF1	C	0.43	0.50	6.58E-05	5.97E-05	220	0.75 (0.65–0.87)
Eosinophilic Esophagitis	446/2586	5	rs4143832	131862977	IL5-IL13	T	0.24	0.18	4.70E-06	1.70E-05	200	1.55 (1.29–1.87)
Eosinophilic Esophagitis	446/2586	5	rs12663750	131971902	IL5-IL13	T	0.28	0.19	3.03E-09	1.00E-06	100	1.73 (1.44–2.07)
Eosinophilic Esophagitis	446/2586	5	rs20541	131995964	IL5-IL13	A	0.26	0.19	3.72E-07	3.00E-06	150	1.61 (1.34–1.94)
Mental retardation	297/1840	5	rs11167764	141479065	NDFIP1	A	0.29	0.20	1.29E-06	4.00E-06	150	1.66 (1.35–2.04)
Mental retardation	297/1840	5	rs77110703	141479833	NDFIP1	T	0.29	0.20	5.83E-07	2.00E-06	150	1.69 (1.38–2.08)
Mental retardation	297/1840	5	rs10057309	141479870	NDFIP1	T	0.29	0.20	4.33E-07	2.00E-06	150	1.70 (1.39–2.09)
Developmental disorders	975/1840	2	rs1595825	198875464	PLCL1	A	0.15	0.21	1.13E-08	2.00E-06	150	0.65 (0.57–0.76)
Supportive otitis media	362/3082	1	rs10801047	191559356	near RGS1	A	0.13	0.08	1.61E-06	2.00E-06	250	1.77 (1.40–2.24)
Depression	107/2864	14	rs7141420	79899454	NRXN3	C	0.66	0.46	4.76E-05	1.10E-04	100	1.78 (1.34–2.34)

\*P (permute): empirical permutation p values after case and control labels are permuted randomly (up to 1,000,000). All results were at the level of FDR- $q < 0.05$ .

\*\*“Cases needed” refers to the estimated number of cases needed to achieve 80% power to detect an association at  $\alpha = 0.05$  given the identified odds ratio and the MAF in this population.



Because of the pleiotropic effects between EoE and other allergy related traits, in addition to conditional analyses, we also found possible synergistic effects. One of the closely related phenotypes with EoE is the presence of food allergy. When we combined these two as a subgroup, two additional effects were identified. One cluster was in *IL1RL1* that was previously associated with the related phenotype, i.e., allergy and asthma (best SNP rs3771180,  $p = 5.71 \times 10^{-5}$ , **Table 2**, Torgerson et al., 2011) and another was in *CLEC16A*, previously associated with different autoimmune diseases [best SNP rs12924729,  $p = 3.34 \times 10^{-8}$  (**Table 2**), (Mells et al., 2011)] and was reported as a suggestive effect in recent GWAS study for EoE (Kottyan et al., 2014).

Variants near *RGS* cluster of genes on chromosome 1, previously reported to be associated with IBD and other autoimmune diseases (Hunt et al., 2008; Esposito et al., 2010), were associated with susceptibility to infection, in particular suppurative otitis media [best SNP rs10801047,  $p = 1.61 \times 10^{-6}$ ,  $p_{\text{perm}} = 2.00 \times 10^{-6}$ ,  $OR = 1.77$  95%CI = 1.398 – 2.24].

New association signals have been detected near the *NDFIP1* gene for mental retardation related traits. Variants near this gene that is expressed mostly in brain, were previously reported to be associated with IBD through an unknown mechanism and with a risk effect for major allele (SNP = rs11167764) (Franke et al., 2010). Instead, we found a risk effect for the minor allele [best SNP rs10057309,  $p = 4.33 \times 10^{-7}$ ,  $p_{\text{perm}} = 2.00 \times 10^{-6}$ ,  $OR = 1.702$ , 95%CI = 1.38 – 2.09] (**Table 3**). Similarly, cerebral palsy, which is linked to mental retardation, was also associated with this variant ( $p = 9.00 \times 10^{-4}$ ). However, conditional analyses controlling for cerebral palsy suggest an independent effect for overall mental retardation (conditional  $p = 8.00 \times 10^{-4}$ ). Furthermore, excluding the small number of samples with known chromosomal abnormalities ( $N < 40$ ) did not affect this result. The overall cluster effect in this neighborhood for mental retardation bolsters the suspicion that an association is found here (**Figure 2C**).

Additionally, for developmental delays of speech and language, a novel signal effect was detected in the *PLCL1* gene at chromosome 2 [best SNP rs1595825,  $p = 1.13 \times 10^{-8}$ ,  $OR = 0.65$  (0.57 – 0.76)] (**Figure 2D**, **Table 3**). Weaker associations ( $0.01 > p > 0.00001$ ) were also detected for related neurologic phenotypes including abnormal movement, lack of coordination and epilepsy at this locus (data not shown).

*NRXN3* polymorphisms that have been previously reported to be associated with substance dependence (Docampo et al., 2012), smoking behavior and attention related problems (Stoltenberg et al., 2011), were associated with depression in our pediatric cohorts (**Table 3**). Noteworthy, the major allele of our reported SNP (rs7141420) has been linked to obesity in adult cohorts (Berndt et al., 2013), while we found association with the minor allele for depression [ $p = 4.76 \times 10^{-5}$ ,  $OR = 1.78$  (1.34 – 2.34), **Table 3**]. Furthermore, rare micro-deletions in this gene were previously reported for Autism case reports but these rare variants are not available to assess in our genotyped cohorts (Vaags et al., 2012).

## DISCUSSION

This first pediatric PheWAS finds 38 associations, 24 previously known phenotype-genotype associations in a pediatric

population using EMR-linked eMERGE databases and identified 14 new possible associations at  $\beta > 0.8$  and  $FDR-q < 0.05$ . From analysis performed on EMR-linked data from 4268 European individuals, we successfully confirmed several major effects for phenotypes with moderate to large sample size, in particular for Asthma, Autism, and neurodevelopmental disease as well as several effects for Type 1 and Type 2 Diabetes (T1DM, T2DM) and Thyroiditis. Almost all of the significant phenotype associations were with common variants ( $MAF > 10\%$ ) (**Tables 2, 3**). In addition, we compared and verified the consistency of allele frequency of reported markers among cohorts, sample collection sites and with CEU-Hapmap data. Considering a desired power of 0.8, for variants at the fixed allele frequency of 10% and size effect of 1.5 or above, 200 cases are sufficient to detect association at an alpha level of 0.05. Indeed, we have surpassed this level for most of our reported traits. In addition, for all reported phenotypes the control sample was at least two or three times larger than cases (**Tables 2, 3**). Importantly, since our control samples for each trait are an EMR-derived population and not healthy individuals, this large number of control samples provides minor allele frequencies consistent with hapmap-CEU frequencies for all of our reported variants.

The results for JRA and EoE depend upon previously published studies of these phenotypes. While the case samples are mostly identical, the control samples were substantially different. Consequently, we cannot refer to these particular findings as constituting confirmation and yet our results and different methodology support the previous reports.

In addition, we also identified several novel PheWAS findings for pediatric traits in particular for Allergic Rhinitis, Otitis Media, EoE, Mental Retardation, and Developmental Delays all with sufficient power ( $\beta > 0.8$ ) (**Table 3**, **Figures 2B–D**). This study, however, is underpowered to make discoveries for rare variants or uncommon traits. The power to detect a finding in PheWAS is determined by many factors, including sample size, risk allele frequency, effect size, model of inheritance, the effect of environment and the prevalence of a phenotype within the population.

Similar to previous studies, we also observed pleiotropy for a number of loci in particular *PTPN22* for JRA, T1DM, and Thyroiditis, *IL5* for Eosinophilia, Asthma, and EoE and *NDFIP1* for Mental Retardation traits and Cerebral Palsy. These pleiotropic effects are specifically expected to be due to underlying biologic correlations. On the other hand, we rarely observed simultaneous robust associations with multiple unrelated phenotypes that had sufficient power. Furthermore, one of the advantages of PheWAS studies is the ability to control the granularity of a database with regard to related phenotypes. For example, by combining two related phenotypes such as uveitis with JRA or food allergy with EoE, we were able to evaluate new subgroups and identify new loci responsible for shared underlying pathways that otherwise cannot be detected or require much larger sample sizes. Further studies with larger sample sizes would be useful to test and perhaps corroborate these findings.

Association of Allergic Rhinitis with loci responsible for diabetes in adults (*GCKR-JAZF1*) may highlight a shared underlying mechanism. In fact, the connection between allergy and diabetes

has been previously suggested in humans but cannot be explained by the Th1/Th2 paradigm (Dales et al., 2005). Moreover, in animal experiments, treating mice with mast cell-stabilizing agents reduced diabetes manifestations (Liu et al., 2009). It is also possible that in our pediatric cohorts we have under-diagnosed children who are diagnosed with diabetes which would appear in a later stage of development. In fact, *GCKR* is an inhibitor of glucokinase (*GCK*), a gene responsible for the autosomal dominant form of T2DM that usually develops later in life and in adulthood. Of note, neither of these two loci showed significant association with Body Mass Index (BMI) in our previous report with these data nor has the obesity link been established in adult studies (Namjou et al., 2013).

The novel association of a cytokine cluster in the *IL5-IL13* region for the EoE trait is particularly interesting since anti-IL5 monoclonal antibodies have been recommended as a novel therapeutic agent for EoE and other eosinophilia-related traits (Corren, 2012). In general, both *IL5* and *IL13* play a major role for regulation of maturation, recruitment, and survival of eosinophils and the variant reported here has been previously associated with other allergic-related traits and with the same direction of allele frequency difference (Bottema et al., 2008; Granada et al., 2012). In particular, a non-synonymous polymorphism in the *IL13* gene, rs20541 (R130Q) (Table 3), has been shown to be associated with increased IL-13 protein activity, altered IL-13 production, and increased binding of nuclear proteins to this region (van der Pouw Kraan et al., 1999). Perhaps, the association is a reflection of linkage disequilibrium with another polymorphism in the 5q31 region. In fact, in our analyses residual effect still exists for the best SNP (rs12653750), shown in Figure 2B after controlling for rs20541 ( $p$ -conditional =  $2.27 \times 10^{-5}$ ) ( $r^2 = 0.35$ ). This possible association did not reach significance in previous GWAS studies for EoE and had only produced a suggestive effect ( $0.05 < p < 0.001$ ). Perhaps, this behavior is explained partly by phenotypic heterogeneity since minor allele frequency of independent set of both control populations were the same. Indeed, we found that those with the subphenotype of EoE with Eosinophilia had the strongest size effect ( $OR = 1.83$ , 95%CI =  $1.44 - 2.32$ ) and our cohorts were enriched with this subphenotype [177 of total 446 EoE cases (40%)]. Of note, the SNPs in this region were originally selected because of eosinophilia-related publications (Bottema et al., 2008; Granada et al., 2012).

Moreover, combining subgroups of patients with food allergy and EoE revealed two new loci that may explain shared etiology. Indeed, the connection between allergy and Interleukin 1 receptor-like-1 (*IL1R1*) is already known (Torgerson et al., 2011). The ligand for *IL1R1*, IL-33, is a potent eosinophil activator (Bouffi et al., 2013). Interestingly, there is also a report of association of *CLEC16A* variants with allergy in large analysis with more than 50,000 subjects from 23andMe Inc. (Hinds et al., 2013). C-type lectin domain family 16, also known as *CLEC16A*, is mostly associated with autoimmune related traits and is highly expressed in B lymphocytes and natural killer cells. The molecular and cellular functions of *CLEC16A* are currently under investigation.

Our conditional analyses suggest an independent effect at the *SLC22A4* gene for Atopic Dermatitis. This solute carrier family

gene is predominantly expressed in CD14 cells and has an important role for elimination of many endogenous small organic cations as well as a wide array of drugs and environmental toxins. The associated SNP, rs272889, has been previously shown to be correlated with blood metabolite concentration (Suhre et al., 2011). Other variants in this gene were associated with IBD and Crohns disease as well (Feng et al., 2009). Of note, a key substrate of this transporter is ergothioneine, a natural antioxidant, which Mammalia acquire exclusively from their food. Ergothioneine is a powerful antioxidant though its precise physiological purpose remains unclear.

Asthma is associated at the 17q21 in our cohorts (Figure 1). The best associated SNP, rs8067378, is known to function as a cis-regulatory variant that correlates with expression of the *GSDMB* gene (Verlaan et al., 2009). Variants in *GSDMB* have been shown to determine multiple asthma related phenotypes specifically in childhood asthma including associations with lung function and disease severity (Tulah et al., 2013). These gasdermin-family genes are implicated in the regulation of apoptosis mostly in epithelial cells and have also been linked to cancer; however, their actual function with respect to disease association remains unknown. The associated variants in this cluster are suspected to be regulatory SNPs that govern the transcriptional activity of at least three nearby genes (*ZBP2*, *GSDMB*, and *ORMDL3*) (Verlaan et al., 2009).

We confirmed several loci responsible for Autism and Pervasive Developmental Disease including *MACROD2*, *ITGB3*, *CADM2*, and *GRIK2*. *ITGB3* has been known as a quantitative trait locus (QTL) for whole blood serotonin levels (Weiss et al., 2004, 2006). Serotonin is a monoamine neurotransmitter that has long been implicated in the etiology of Autism. In fact, about 30 percent of patients with autism have abnormal blood serotonin levels (Weiss et al., 2004). Similarly, *GRIK2* is an ionotropic glutamate receptor associated with autism (Cook, 1990; Cook et al., 1997). *CADM2* is a member of the synaptic cell adhesion molecule with roles in early postnatal development of the central nervous system (Thomas et al., 2008). The function of *MACROD2* (previously c20orf133) is still largely unknown. For Autism that is more commonly seen in males, we found no significant gender effect for these loci.

Association of variants in the neighborhood of RGS cluster genes with suppurative otitis media is another novel finding. SNPs in this region have been previously linked to celiac disease, multiple sclerosis and other autoimmune diseases (Hunt et al., 2008; Esposito et al., 2010). The link between susceptibility to infection and autoimmunity has been long suggested given the fact that the level and regulation of RGS proteins in lymphocytes also significantly impact lymphocyte migration and function. In our pediatric cohort the number of patients with celiac disease was small ( $n = 23$ ) and the association was not detected. Interestingly, one of the major risk variant for celiac disease, rs13151961 (KIAA1109), as well as known HLA variants, produced a trend toward association for celiac disease but did not pass the FDR threshold (data not shown).

Finally we also detected a novel association between mental retardation and the *NDFIP1* gene (Figure 2C, Table 3). Of note, no effect was detected with Autism at this locus. Indeed, the

only other effect observed in this region was related to Cerebral Palsy ( $p = 9.00 \times 10^{-4}$ ) and, as mentioned above, an independent effect exists for Mental Retardation. The PheWAS code for mental retardation includes ICD-9 codes for mild, moderate and profound degrees of retardation as well as not-otherwise-specified (MR-NOS). Indeed, an additive correlation can also be detected when we score these subgroups according to severity excluding the MR-NOS subgroup ( $p = 3.00 \times 10^{-4}$ ). Larger sample size is necessary to fully elucidate this interesting effect. The Nedd4 family-interacting protein 1 (Ndfip1) is an adaptor protein for the Nedd4 family of E3 ubiquitin ligases important for axon and dendrite development. In fact, cerebral atrophy is one of the main findings in Ndfip1 KO mice (Hammond et al., 2014). Another neurodevelopmental association effect was observed in the vicinity of the Phospholipase C-Like 1 (PLCL1, PRIP-1) gene for overall Developmental Delays-Speech and Language Disorder (Table 3, Figure 2D). This gene which is expressed predominantly in brain, regulates the turnover of GABA-receptors, contributes to the maintenance of GABA-mediated synaptic inhibition, and has been implicated in several pathologies in animal models and human including epilepsy, bone density and cancer (Liu et al., 2008; Zhu et al., 2012). Finally, we also detected a link between Neuroxin-3 and early onset depression in this study (Table 3). In fact, this gene has a major role in synaptic plasticity and function in the nervous system as a receptor and cell adhesion molecule.

In summary, by using the PheWAS approach and re-mapping the ICD-9 codes on our European ancestry pediatric cohorts we have been able to verify and confirm a variety of previously reported associations as well as discover new effects that potentially have clinical implications. Similar to adult PheWAS studies, our data also support the importance of this approach in pediatrics. We replicated known phenotype-genotype associations in a pediatric population using these EMR-linked eMERGE databases, and also noted a number of new possible associations that warrant additional study, especially including the relationship of *PLCL1* to speech and language development and *IL5-IL13* to EoE. Some of the limitations to the current PheWAS map include the fact that current map does not take into account of the correlation between some phenotypes and treat them as independent. Future pediatric PheWAS directions will include enhancements of a PheWAS map for more precise modeling of trait associations as well as improvements for richer querying and filtering.

## ACKNOWLEDGMENTS

We are grateful to the individuals who participated in this study. We thank the genotyping core facilities in both academic centers (CCHMC, BCH) and our colleagues who facilitated the genotyping and recruitment of subjects.

This work was supported by a grant from the National Human Genomic Research Institute: 1U01HG006828 with other NIH support (R37 AI024717, P01 AI083194, U19 AI066738, and P01 AR049084), the US Department of Veterans Affairs, the Campaign Urging Research For Eosinophilic Diseases (CURED) Foundation, as well as the Food Allergy Research Education (FARE) Foundation.

## REFERENCES

- Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., et al. (2010). A genome-wide scan for common alleles affecting risk for autism. *Hum. Mol. Genet.* 19, 4072–4082. doi: 10.1093/hmg/ddq307
- Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* 41, 703–707. doi: 10.1038/ng.381
- Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* 45, 501–512. doi: 10.1038/ng.2606
- Bi, M., Kao, W. H., Boerwinkle, E., Hoogveen, R. C., Rasmussen-Torvik, L. J., Astor, B. C., et al. (2010). Association of rs780094 in GCKR with metabolic traits and incident diabetes and cardiovascular disease: the ARIC Study. *PLoS ONE* 5:e11690. doi: 10.1371/journal.pone.0011690
- Bottema, R. W., Reijmerink, N. E., Kerkhof, M., Koppelman, G. H., Stelma, F. F., Gerritsen, J., et al. (2008). Interleukin 13, CD14, pet and tobacco smoke influence atopy in three Dutch cohorts: the allergenic study. *Eur. Respir. J.* 32, 593–602. doi: 10.1183/09031936.00162407
- Bouffi, C. Ist., Rochman, M., Züst, C. B., Stucke, E. M., Kartashov, A., Fulkerson, P. C., et al. (2013). IL-33 markedly activates murine eosinophils by an NF- $\kappa$ B-dependent mechanism differentially dependent upon an IL-4-driven auto-inflammatory loop. *J. Immunol.* 91, 4317–4325. doi: 10.4049/jimmunol.1301465
- Carroll, R. J., Bastarache, L., and Denny, J. C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 30, 2375–2376. doi: 10.1093/bioinformatics/btu197
- Cook, E. H. Jr., Courchesne, R., Lord, C., Cox, N. J., Yan, S., Lincoln, A., et al. (1997). Evidence of linkage between the serotonin transporter and autistic disorder. *Mol. Psychiatry* 2, 247–250.
- Cook, E. H. (1990). Autism: review of neurochemical investigation. *Synapse* 6, 292–308. doi: 10.1002/syn.890060309
- Corren, J. (2012). Inhibition of interleukin-5 for the treatment of eosinophilic diseases. *Discov. Med.* 13, 305–312.
- Dales, R., Chen, Y., Lin, M., and Karsh, J. (2005). The association between allergy and diabetes in the Canadian population: implications for the Th1-Th2 hypothesis. *Eur. J. Epidemiol.* 20, 713–717. doi: 10.1007/s10654-005-7920-1
- Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31, 1102–1110. doi: 10.1038/nbt.2749
- Denny, J. C., Crawford, D. C., Ritchie, M. D., Bielinski, S. J., Basford, M. A., Bradford, Y., et al. (2011). Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet.* 89, 529–542. doi: 10.1016/j.ajhg.2011.09.008
- Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26, 1205–1210. doi: 10.1093/bioinformatics/btq126
- Docampo, E., Ribasés, M., Gratacòs, M., Bruguera, E., Cabezas, C., Sánchez-Mora, C., et al. (2012). Association of Neurexin 3 polymorphisms with smoking behavior. *Genes Brain Behav.* 11, 704–711. doi: 10.1111/j.1601-183X.2012.00815.x
- Esposito, F., Patsopoulos, N. A., Cepok, S., Kockum, I., Leppä, V., Booth, D. R., et al. (2010). IL12A, MPHOSPH9/CDK2AP1 and RGS1 are novel multiple sclerosis susceptibility loci. *Genes Immun.* 11, 397–405. doi: 10.1038/gene.2010.28
- Feng, Y., Zheng, P., Zhao, H., and Wu, K. (2009). SLC22A4 and SLC22A5 gene polymorphisms and Crohn's disease in the Chinese Han population. *J. Dig. Dis.* 10, 181–187. doi: 10.1111/j.1751-2980.2009.00383.x
- Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 12, 1118–1125. doi: 10.1038/ng.717
- Gauderman, W. J., and Morrison, J. M. (2006). *QUANTO 1.1: A Computer Program for Power and Sample Size Calculations for Genetic-epidemiology Studies*. Available online at: <http://hydra.usc.edu/gxe>
- Gidding, S. S. (1993). The rationale for lowering serum cholesterol levels in American children. *Am. J. Dis. Child.* 147, 386–392.

- Granada, M., Wilk, J. B., Tuzova, M., Strachan, D. P., Weidinger, S., Albrecht, E., et al. (2012). A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. *J. Allergy Clin. Immunol.* 129, 840–845.e21. doi: 10.1016/j.jaci.2011.09.029
- Hammond, V. E., Gunnarsen, J. M., Goh, C. P., Low, L. H., Hyakumura, T., Tang, M. M., et al. (2014). Ndfip1 is required for the development of pyramidal neuron dendrites and spines in the neocortex. *Cereb. Cortex* 24, 3289–3300. doi: 10.1093/cercor/bht191
- Hinds, D. A., McMahon, G., Kiefer, A. K., Do, C. B., Eriksson, N., Evans, D. M., et al. (2013). A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. *Nat. Genet.* 45, 907–911. doi: 10.1038/ng.2686
- Hinks, A., Cobb, J., Marion, M. C., Prahalad, S., Sudman, M., Bowes, J., et al. (2013). Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet.* 45, 664–669. doi: 10.1038/ng.2614
- Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)* 1, 457–470. doi: 10.1534/g3.111.001198
- Huertas-Vazquez, A., Plaisier, C., Weissglas-Volkov, D., Sinsheimer, J., Canizales-Quinteros, S., Cruz-Bautista, I., et al. (2008). TCF7L2 is associated with high serum triacylglycerol and differentially expressed in adipose tissue in families with familial combined hyperlipidaemia. *Diabetologia* 51, 62–69. doi: 10.1007/s00125-007-0850-6
- Hunt, K. A., Zhernakova, A., Turner, G., Heap, G. A., Franke, L., Bruinenberg, M., et al. (2008). Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* 40, 395–402. doi: 10.1038/ng.102
- Jamain, S., Betancur, C., Quach, H., Philippe, A., Fellous, M., Giros, B., et al. (2002). Linkage and association of the glutamate receptor 6 gene with autism. *Mol. Psychiatry* 7, 302–310. doi: 10.1038/sj.mp.4000979
- Klein, J. D., Sesselberg, T. S., Johnson, M. S., O'Connor, K. G., Cook, S., Coon, M., et al. (2010). Adoption of body mass index guidelines for screening and counseling in pediatric practice. *Pediatrics* 125, 265–272. doi: 10.1542/peds.2008-2985
- Kottyan, L. C., Davis, B., Sherrill, J. D., Liu, K., Rochman, M., Kaufman, K., et al. (2014). Identification of genome-wide susceptibility loci for eosinophilic esophagitis elucidates tissue-specificity of this allergic disease. *Nat. Genet.* 46, 895–900. doi: 10.1038/ng.3033
- Lee, H. S., Kang, J., Yang, S., Kim, D., and Park, Y. (2011). Susceptibility influence of a PTPN22 haplotype with thyroid autoimmunity in Koreans. *Diabetes Metab. Res. Rev.* 27, 878–882. doi: 10.1002/dmrr.1265
- Li, Y., Mao, Q., Shen, L., Tian, Y., Yu, C., Zhu, W. M., et al. (2010). Interleukin-23 receptor genetic polymorphisms and Crohn's disease susceptibility: a meta-analysis. *Inflamm. Res.* 59, 607–614. doi: 10.1007/s00011-010-0171-y
- Liao, K. P., Diogo, D., Cui, J., Cai, T., Okada, Y., Gainer, V. S., et al. (2014). Association between low density lipoprotein and rheumatoid arthritis genetic factors with low density lipoprotein levels in rheumatoid arthritis and non-rheumatoid arthritis controls. *Ann. Rheum. Dis.* 73, 1170–1175. doi: 10.1136/annrheumdis-2012-203202
- Liu, J., Divoux, A., Sun, J., Zhang, J., Clément, K., Glickman, J. N., Sukhova, G. K., et al. (2009). Genetic deficiency and pharmacological stabilization of mast cells reduce diet-induced obesity and diabetes in mice. *Nat. Med.* 15, 940–945. doi: 10.1038/nm.1994
- Liu, Y. Z., Wilson, S. G., Wang, L., Liu, X. G., Guo, Y. F., Li, J., et al. (2008). Identification of PLCL1 gene for hip bone size variation in females in a genome-wide association study. *PLoS ONE* 3:e3160. doi: 10.1371/journal.pone.0003160
- Lyssenko, V., Lupi, R., Marchetti, P., Del Guerra, S., Orho-Melander, M., Almgren, P., et al. (2007). Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J. Clin. Invest.* 117, 2155–2163. doi: 10.1172/JCI30706
- Mells, G. F., Floyd, J. A., Morley, K. I., Cordell, H. J., Franklin, C. S., Shin, S. Y., et al. (2011). Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* 43, 329–332. doi: 10.1038/ng.789
- Namjou, B., Keddache, M., Marsolo, K., Wagner, M., Lingren, T., Cobb, B., et al. (2013). EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children. *Front. Genet.* 4:268. doi: 10.3389/fgene.2013.00268
- Namjou, B., Sestak, A. L., Armstrong, D. L., Zidovetzki, R., Kelly, J. A., Jacob, N., et al. (2009). High-density genotyping of STAT4 reveals multiple haplotypic associations with systemic lupus erythematosus in different racial groups. *Arthritis Rheum.* 60, 1085–1095. doi: 10.1002/art.24387
- Neuraz, A., Chouchana, L., Malamut, G., Le Beller, C., Roche, D., Beaune, P., et al. (2013). Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS Comput. Biol.* 9:e1003405. doi: 10.1371/journal.pcbi.1003405
- Onuma, H., Tabara, Y., Kawamoto, R., Shimizu, I., Kawamura, R., Takata, Y., et al. (2010). The GCKR rs780094 polymorphism is associated with susceptibility of type 2 diabetes, reduced fasting plasma glucose levels, increased triglycerides levels and lower HOMA-IR in Japanese population. *J. Hum. Genet.* 55, 600–604. doi: 10.1038/jhg.2010.75
- Paternoster, L., Standl, M., Chen, C. M., Ramasamy, A., Bønnelykke, K., Duijts, L., et al. (2011). Meta-analysis of genome-wide association studies identifies three new risk loci for atopic dermatitis. *Nat. Genet.* 44, 187–192. doi: 10.1038/ng.1017
- Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., et al. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet.* 9:e1003087. doi: 10.1371/journal.pgen.1003087
- Pendergrass, S. A., Brown-Gentry, K., Dudek, S. M., Torstenson, E. S., Ambite, J. L., Avery, C. L., et al. (2011). The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* 35, 410–422. doi: 10.1002/gepi.20589
- Plenge, R. M., Seielstad, M., Padyukov, L., Lee, A. T., Remmers, E. F., Ding, B., et al. (2007). TRAF1-C5 as a risk locus for rheumatoid arthritis—a genome-wide study. *N. Engl. J. Med.* 357, 1199–1209. doi: 10.1056/NEJMoa073491
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847
- Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., et al. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–2337. doi: 10.1093/bioinformatics/btq419
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795
- Pykalainen, M., Kinos, R., Valkonen, S., Rydman, P., Kilpeläinen, M., Laitinen, L. A., et al. (2005). Association analysis of common variants of STAT6, GATA3, and STAT4 to asthma and high serum IgE phenotypes. *J. Allergy Clin. Immunol.* 115, 80–87. doi: 10.1016/j.jaci.2004.10.006
- Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., et al. (2013). Genome and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 127, 1377–1385. doi: 10.1161/CIRCULATIONAHA.112.000604
- Rothenberg, M. E., Spergel, J. M., Sherrill, J. D., Annaiah, K., Martin, L. J., Cianferoni, A., et al. (2010). Common variants at 5q22 associate with pediatric eosinophilic esophagitis. *Nat. Genet.* 42, 289–291. doi: 10.1038/ng.547
- Setia, S., Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou, B., et al. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* 5:370. doi: 10.3389/fgene.2014.00370
- Stoltenberg, S. F., Lehmann, M. K., Christ, C. C., Hersrud, S. L., and Davies, G. E. (2011). Associations among types of impulsivity, substance use problems and neurexin-3 polymorphisms. *Drug Alcohol Depend.* 119, e31–e38. doi: 10.1016/j.drugalcdep.2011.05.025
- St Pourcain, B., Skuse, D. H., Mandy, W. P., Wang, K., Hakonarson, H., Timpson, N. J., et al. (2014). Variability in the common genetic architecture of social-communication spectrum phenotypes during childhood and adolescence. *Mol. Autism* 5:18. doi: 10.1186/2040-2392-5-18
- Suhre, K., Shin, S. Y., Petersen, A. K., Mohny, R. P., Meredith, D., Wägele, B., et al. (2011). Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 477, 54–60. doi: 10.1038/nature10354
- Thomas, L. A., Akins, M. R., and Biederer, T. (2008). Expression and adhesion profiles of SynCAM molecules indicate distinct neuronal functions. *J. Comp. Neurol.* 510, 47–67. doi: 10.1002/cne.21773
- Thompson, S. D., Marion, M. C., Sudman, M., Ryan, M., Tsoras, M., Howard, T. D., et al. (2012). Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. *Arthritis Rheum.* 64, 2781–2791. doi: 10.1002/art.34429
- Tiisala, R., and Kantero, R. L. (1971). Studies on growth of Finnish children from birth to 10 years. 3. Comparison of height and weight distance curves based

- on longitudinal and cross-sectional series from birth to 10 years. *Acta Paediatr Scand. Suppl.* 220, 13–7.
- Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., et al. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* 39, 857–864. doi: 10.1038/ng2068
- Torgerson, D. G., Ampleford, E. J., Chiu, G. Y., Gauderman, W. J., Gignoux, C. R., Graves, P. E., et al. (2011). Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.* 43, 887–892. doi: 10.1038/ng.888
- Tulah, A. S., Holloway, J. W., and Sayers, I. (2013). Defining the contribution of SNPs identified in asthma GWAS to clinical variables in asthmatic children. *BMC Med. Genet.* 14:100. doi: 10.1186/1471-2350-14-100
- Vaags, A. K., Lionel, A. C., Sato, D., Goodenberger, M., Stein, Q. P., Curran, S., et al. (2012). Rare deletions at the neurexin 3 locus in autism spectrum disorder. *Am. J. Hum. Genet.* 90, 133–141. doi: 10.1016/j.ajhg.2011.11.025
- van der Pouw Kraan, T. C., van Veen, A., Boeijs, L. C., van Tuyl, S. A., de Groot, E. R., Stapel, S. O., et al. (1999). An IL-13 promoter polymorphism associated with increased risk of allergic asthma. *Genes Immun.* 1, 61–65. doi: 10.1038/sj.gene.6363630
- Verlaan, D. J., Berlivet, S., Hunninghake, G. M., Madore, A. M., Larivière, M., Moussette, S., et al. (2009). Allele-specific chromatin remodeling in the ZBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *Am. J. Hum. Genet.* 85, 377–393. doi: 10.1016/j.ajhg.2009.08.007
- Weiss, L. A., Kosova, G., Delahanty, R. J., Jiang, L., Cook, E. H., Ober, C., et al. (2006). Variation in ITGB3 is associated with whole-blood serotonin level and autism susceptibility. *Eur. J. Hum. Genet.* 14:923–931. doi: 10.1038/sj.ejhg.5201644
- Weiss, L. A., Veenstra-Vanderweele, J., Newman, D. L., et al. (2004). Genomewide association study identifies ITGB3 as a QTL for whole blood serotonin. *Eur. J. Hum. Genet.* 12, 949–954. doi: 10.1038/sj.ejhg.5201239
- Zhu, G., Yoshida, S., Migita, K., Yamada, J., Mori, F., Tomiyama, M., et al. (2012). Dysfunction of extrasynaptic GABAergic transmission in phospholipase C-related, but catalytically inactive protein 1 knockout mice is associated with an epilepsy phenotype. *J. Pharmacol. Exp. Ther.* 340, 520–528. doi: 10.1124/jpet.111.182386
- Conflict of Interest Statement:** The Guest Associate Editor Mariza De Andrade declares that, despite having collaborated with authors Bahram Namjou, Joshua C. Denny, Leah C. Kottyan, Marylyn D. Ritchie, and Shefali S. Verma, the review process was handled objectively and no conflict of interest exists. The Review Editor Andrew Skol declares that, despite having collaborated with author John B. Harley, the review process was handled objectively and no conflict of interest exists. Marc E. Rothenberg is a consultant for Immune Pharmaceuticals and has an equity interest. Marc E. Rothenberg has a royalty interest in reslizumab being developed by Teva Pharmaceuticals. Marc E. Rothenberg, John B. Harley, and Leah C. Kottyan are co-inventors of a patent application, being submitted by CCHMC, concerning the genetics of EoE. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 May 2014; accepted: 31 October 2014; published online: 18 November 2014.

Citation: Namjou B, Marsolo K, Carroll RJ, Denny JC, Ritchie MD, Verma SS, Lingren T, Porollo A, Cobb BL, Perry C, Kottyan LC, Rothenberg ME, Thompson SD, Holm IA, Kohane IS and Harley JB (2014) Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. *Front. Genet.* 5:401. doi: 10.3389/fgene.2014.00401

This article was submitted to *Applied Genetic Epidemiology*, a section of the journal *Frontiers in Genetics*.

Copyright © 2014 Namjou, Marsolo, Carroll, Denny, Ritchie, Verma, Lingren, Porollo, Cobb, Perry, Kottyan, Rothenberg, Thompson, Holm, Kohane and Harley. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.